On Approaches for Monitoring Categorical Event Series

Helmut Schmidt Universität
Universität der Bundeswehr Hamburg

MATH
STAT

Christian H. Weiß

Department of Mathematics & Statistics,
Helmut Schmidt University, Hamburg

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH
STAT

Monitoring Categorical Event Series

Introduction & Outline

Statistical process control (SPC): (Montgomery, 2009)
monitor quality-related processes, for example,
in manufacturing, service industries, health surveillance.

Control chart: certain statistics computed sequentially
in time and used to decide about actual state of process.

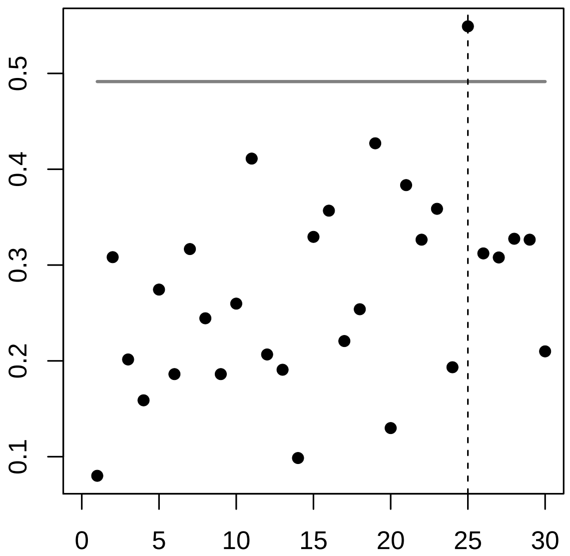
No intervention in process if **in control**, i. e., if monitored
statistics stationary according to specified time series model
(e. g., independent and identically distributed (i. i. d.)
with specified marginal distribution).

By contrast, if deviations from in-control model, such as shifts or drifts in model parameters, then process **out of control**.

In traditional control chart applications, we compare plotted statistics against given control limits. If statistic beyond limits, then alarm triggered to indicate possible out-of-control situation.

Example control chart:
(see below)

Alarm at $t = 25$, because upper control limit violated.



Aim: true alarm as soon as possible,

but avoid false alarm for as long as possible.

Here, waiting time until first alarm is

run length of control chart.

Should be large (low) if process in control (out of control).

For these and further basics on SPC and control charts,

see textbook by Montgomery (2009).

Most SPC literature about quality characteristics measured

on continuous quantitative scale (**variables charts**).

Here: discrete-valued characteristics, **attributes charts**.

Focus on **qualitative data** monitored sequentially in time,

thus **categorical event series** $(X_t)_{t\in\mathbb{N}=\{1,2,...\}}$.

Quality features X_t have finite range $\mathcal{S}$ of either

- unordered but distinguishable categories (**nominal data**)
- or categories exhibiting natural order (**ordinal data**).

Unique notation: $\mathcal{S} = \{s_0, s_1, \ldots, s_d\}$ with some $d \in \mathbb{N}$,

arranged in lexicographical order (nominal case)

or natural order (ordinal case).

Special case $d = 1$ (X_t **binary**): 0–1 coding $s_0 := 0$, $s_1 := 1$.

Many manufacturing applications, for example:

- six nominal types of paint defect on manufactured ceiling fan covers (Mukhopadhyay, 2008);
- four ordinal categories of flash on head of electric toothbrushes (Li et al., 2014).

Further examples & references in **main paper for this talk**,

Weiß (2021) On approaches for monitoring categorical event series. In K.P. Tran (ed.): *Control Charts and Machine Learning for Anomaly Detection in Manufacturing*, Springer Series in Reliability Engineering; 105–129.

In Weiß (2021), survey of **three approaches** for monitoring categorical event series:

1. Monitor **counts of events** for successive time intervals, i. e., categorical data transformed into count time series, use control charts for count data.
2. **Control charts** directly for categorical event series. Different solutions required for nominal vs. ordinal data.
3. **Rule-based machine learning procedures** to categorical event series (scalability w. r.t. amount/complexity of data).

Because of time limitations: sketches for Approaches 1 & 3, detailed discussion and data examples for **Approach 2**.

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Monitoring Time Series of Event Counts

Sketch

Different ways of transforming categorical event series $(X_t)_{\mathbb{N}}$ into **count process** $(Y_t)_{\mathbb{N}}$:

- count categorical events (e. g., malfunctions) within fixed time intervals $\Rightarrow$ counts Y_t might become arbitrarily large, i. e., unbounded counts with range $\mathbb{N}_0 = \{0, 1, \ldots\}$;
- count non-conforming items in samples of size $n \in \mathbb{N}$ $\Rightarrow$ bounded counts Y_t with range $\{0, \ldots, n\}$;
- determine discrete waiting time until certain event happens (e. g., number of manufactured items until next defective) $\Rightarrow$ unbounded counts Y_t with range $\mathbb{N}$ or $\mathbb{N}_0$.

Stochastic properties of $(Y_t)_{\mathbb{N}}$ implied by $(X_t)_{\mathbb{N}}$.

Example: binary process $(X_t)_{\mathbb{N}}$ (“defect — yes or no”),
take segments of length $n \in \mathbb{N}$ from $(X_t)_{\mathbb{N}}$
at (distant) inspection times $t_1, t_2, \ldots \in \mathbb{N}$,
event counts per segment, i. e., $Y_r = X_{t_r} + \ldots + X_{t_r+n-1}$.

- If $(X_t)_{\mathbb{N}}$ i. i. d., then $(Y_r)_{\mathbb{N}}$ i. i. d., **binomial distribution**.
- If $(X_t)_{\mathbb{N}}$ stationary Markov chain, i. e.,

$$P(X_t = x \mid X_{t-1} = x_1, X_{t-2} = x_2, \ldots) \;=\; p(x|x_1),$$

then $Y_r = X_{t_r} + \ldots + X_{t_r+n-1}$ **Markov-binomial distr.**

Phase-I analysis: For count time series $y_1, \ldots, y_T$, standard tools from time series *analysis* can be applied, such as time series plot or (partial) autocorrelation function.

But for modeling underlying count process $(Y_t)_{\mathbb{N}}$, *tailor-made models* required, such as

- integer-valued autoregressive moving-average (INARMA),
- int.-val. generalized AR conditional heterosced. (INGARCH),
- non-linear regression models, and many more.

See introductory textbook by Weiß (2018a) for overview.

Most basic chart for bounded counts: np-chart

(for unbounded counts: c-chart), see Montgomery (2009).

Shewhart-type chart, where counts $Y_1, Y_2, \ldots$ directly plotted on chart and compared to given control limits $0 \leq l < u$.

Alarm triggered for rth count if $Y_r > u$ or $Y_r < l$.

(analogous to control chart plotted before)

Limits l, u such that certain **ARL performance** (average RL).

Here, ARL expresses *mean* waiting time until first alarm.

Choose l, u such that in-control ARL close to given target value.

But: memory-less, insensitive to small deteriorations.

Popular charts with inherent memory:

CUSUM chart (cumulative sum) dating back to Page (1954),

$$\mathrm{C}_0 = c_0, \qquad \mathrm{C}_r = \max\{0,\ Y_r - k + \mathrm{C}_{r-1}\} \quad \text{for } r = 1, 2, \ldots$$

withe reference value $k > 0$ and (upper) control limit $h > 0$;

EWMA chart (exponentially weighted moving-average)

by Roberts (1959), $\mathrm{Z}_r = \lambda \cdot Y_r + (1-\lambda) \cdot \mathrm{Z}_{r-1}$

with smoothing parameter $\lambda \in (0;1]$ and control limits $0 \leq l < u$.

Example: Rakitzis et al. (2017) investigate np- and CUSUM charts for binomial AR(1) process (count Markov chain).

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Monitoring Categorical Event Series

Analysis & Modeling

As data are qualitative, standard tools from time series analysis cannot be used. Instead, tailor-made solutions are required for both analysis and modeling, where ordinal case to be distinguished from nominal one.

- Arithmetic operations not applicable to $\mathcal{S} = \{s_0, s_1, \ldots, s_d\}$, hence moments not defined for $(X_t)_{\mathbb{N}}$, e. g., no mean, variance, ACF for categorical event series.
- If $(X_t)_{\mathbb{N}}$ **ordinal**, at least quantiles and time series plot.
- If $(X_t)_{\mathbb{N}}$ **nominal**, location by mode, rate evolution graph.

See Weiß (2018a) for a discussion.

Nominal range	Ordinal range
Marginal PMF: (probability mass function) $\boldsymbol{p} = (p_0, \ldots, p_d)^\top \in [0;1]^{d+1}$ with $p_i = P(X = s_i)$	Marginal CDF: (cumulative distribution fct.) $\boldsymbol{f} = (f_0, \ldots, f_{d-1})^\top \in [0;1]^{d}$ with $f_i = P(X \le s_i)$
Bivariate lag-h PMF: $p_{ij}(h) = P(X_t = s_i, X_{t-h} = s_j)$	Bivariate lag-h CDF: $f_{ij}(h) = P(X_t \le s_i, X_{t-h} \le s_j)$
Binarization $(\boldsymbol{Y}_t)_{\mathbb{N}}$ with $Y_{t,i} = \mathbb{1}_{\{X_t = s_i\}}$, so $E[\boldsymbol{Y}_t] = \boldsymbol{p}$	Binarization $(\boldsymbol{Z}_t)_{\mathbb{N}}$ with $Z_{t,i} = \mathbb{1}_{\{X_t \le s_i\}}$, so $E[\boldsymbol{Z}_t] = \boldsymbol{f}$
Sample PMF: $\hat{\boldsymbol{p}} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{Y}_t$	Sample CDF: $\hat{\boldsymbol{f}} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{Z}_t$
Bivariate (cumulative) relative frequencies	
$\hat{p}_{ij}(h) = \frac{1}{T-h}\sum_{t=h+1}^T Y_{t,i}\, Y_{t-h,j}$	$\hat{f}_{ij}(h) = \frac{1}{T-h}\sum_{t=h+1}^T Z_{t,i}\, Z_{t-h,j}$

Note that both **binarizations** lead to different range:

$$\boldsymbol{Y}_t \in \left\{ \begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}, \begin{pmatrix}0\\1\\\vdots\\0\end{pmatrix}, \ldots, \begin{pmatrix}0\\\vdots\\1\\0\end{pmatrix}, \begin{pmatrix}0\\\vdots\\0\\1\end{pmatrix} \right\} =: \{\boldsymbol{e}_0, \ldots, \boldsymbol{e}_d\} \subset [0;1]^{d+1}$$

vs.

$$\boldsymbol{Z}_t \in \left\{ \begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}, \begin{pmatrix}0\\1\\\vdots\\1\end{pmatrix}, \ldots, \begin{pmatrix}0\\\vdots\\0\\1\end{pmatrix}, \begin{pmatrix}0\\\vdots\\0\\0\end{pmatrix} \right\} =: \{\boldsymbol{c}_0, \ldots, \boldsymbol{c}_d\} \subset [0;1]^{d}.$$

Dispersion concepts for qualitative random variables:

- Minimal dispersion iff one-point distribution,

  i. e., $\boldsymbol{p} \in \{\boldsymbol{e}_0, \ldots, \boldsymbol{e}_d\}$ or $\boldsymbol{f} \in \{\boldsymbol{c}_0, \ldots, \boldsymbol{c}_d\}$, respectively.
- Maximal nominal dispersion iff uniform, $\boldsymbol{p} = (\frac{1}{d+1}, \ldots, \frac{1}{d+1})^\top$.
- Max. ordinal disp. iff extreme two-point, $\boldsymbol{f} = (\frac{1}{2}, \ldots, \frac{1}{2})^\top$.

Nominal range	**Ordinal range**
Index of qualitative variation: $\mathrm{IQV} = \frac{d+1}{d}\big(1-\sum_{i=0}^{d} p_i^2\big)$	Index of ordinal variation: $\mathrm{IOV} = \frac{4}{d}\sum_{i=0}^{d-1} f_i(1-f_i)$ Ordinal skewness: $\mathrm{skew} = \frac{2}{d}\sum_{i=0}^{d-1} f_i \ - 1$
Nominal Cohen's κ: $\kappa_{\mathrm{nom}}(h) = \dfrac{\sum_{j=0}^{d}\big(p_{jj}(h)-p_j^2\big)}{1-\sum_{i=0}^{d} p_i^2}$	Ordinal Cohen's κ: $\kappa_{\mathrm{ord}}(h) = \dfrac{\sum_{j=0}^{d-1}\big(f_{jj}(h)-f_j^2\big)}{\sum_{i=0}^{d-1} f_i(1-f_i)}$

The signed κ-measures serve as substitutes of ACF, where positive (negative) values express extend of (dis)agreement between X_t and X_{t-h}, see Weiß (2018a, 2020).

Nominal time series models, see Weiß (2018a), include

- higher-order Markov models,
- variable-length Markov models,
- mixture transition distribution models,
- Hidden-Markov models,
- generalized linear models (GLMs),
- discrete ARMA(p, q) models of Jacobs & Lewis (1983):

(. . .)

Nominal time series models, see Weiß (2018a), include

- (. . .)
- **discrete ARMA($\mathbf{p},\mathbf{q}$) models** of Jacobs & Lewis (1983):

  i. i. d. multinomial "random choice" vectors

$$\boldsymbol{D}_t = (\alpha_{t,1},\ldots,\alpha_{t,\mathrm{p}},\beta_{t,0},\ldots,\beta_{t,\mathrm{q}}) \sim \mathrm{Mult}(1;\ \phi_1,\ldots,\phi_{\mathrm{p}},\varphi_0,\ldots,\varphi_{\mathrm{q}}),$$

  i. i. d. categorical innovations $(\epsilon_t)_{\mathbb{Z}}$ with range $\mathcal{S}$,

$$X_t = \alpha_{t,1}\cdot X_{t-1}+\ldots+\alpha_{t,\mathrm{p}}\cdot X_{t-\mathrm{p}} + \beta_{t,0}\cdot\epsilon_t+\ldots+\beta_{t,\mathrm{q}}\cdot\epsilon_{t-\mathrm{q}},$$

  where $0\cdot s := 0$, $1\cdot s := s$, and $s+0 := s$ for each $s\in\mathcal{S}$.

  Leads to *Yule–Walker equations* for κ-measures.

Modeling approaches for ordinal event series include

- rank-count approach (Weiß, 2020),
  i. e., $X_t = s_{I_t}$ with bounded rank count $I_t \in \{0, \ldots, d\}$,
  uses count model for rank-count process $(I_t)_{\mathbb{N}}$;
- latent-variable approach (Agresti, 2010) with real-valued L_t,
  where $X_t = s_j$ iff $L_t \in [\eta_{j-1}; \eta_j)$ (thus $f_j = F_L(\eta_j)$)
  with thresholds $-\infty = \eta_{-1} < \eta_0 < \ldots < \eta_{d-1} < \eta_d = +\infty$;
  logistic (normal) dist. for L_t gives cumulative logit (probit);
- commonly combined with regression approach (GLMs),
  see Höhle (2010), Li et al. (2018).

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Monitoring Categorical Event Series

Control Charts

Continuous process monitoring

(i. e., monitored statistic updated with each new observation):

- **log-LR CUSUM chart** (log-likelihood ratio) based on in-control $\boldsymbol{p}_0$ and anticipated out-of-control $\boldsymbol{p}_1$:

  $$\mathrm{C}_t \;=\; \max\{0,\, \ell R_t + \mathrm{C}_{t-1}\} \qquad \text{with } \ell R_t \;=\; \sum_{j=0}^{d} Y_{t,j}\, \ln\left(\frac{p_{1,j}}{p_{0,j}}\right),$$

  for i. i. d. data by Ryan et al. (2011), Markov processes by Weiß (2018b), categorical logit regression by Höhle (2010);
- Ye et al. (2002): **EWMA** for nominal binarizations $(\boldsymbol{Y}_t)_{\mathbb{N}}$;
- see Weiß (2021) for further references on approaches for continuous process monitoring.

Sample-based process monitoring

(i. e., samples taken from $(X_t)_{\mathbb{N}}$ to compute plotted statistics).

Compute (absolute) sample frequencies

$\boldsymbol{N}_r = \boldsymbol{Y}_{t_r} + \ldots + \boldsymbol{Y}_{t_r+n-1}$ from nominal binarizations $(\boldsymbol{Y}_t)_{\mathbb{N}}$,

or cumulative frequencies

$\boldsymbol{C}_r = \boldsymbol{Z}_{t_r} + \ldots + \boldsymbol{Z}_{t_r+n-1}$ from ordinal binarizations $(\boldsymbol{Z}_t)_{\mathbb{N}}$.

- Most well-known approach: χ^2-**chart** (Duncan, 1950),

$$\mathrm{X}_r^2 \;=\; \sum_{j=0}^{d} \frac{\big(N_{r,j} - n\,p_{0,j}\big)^2}{n\,p_{0,j}},$$

  measures any kind of deviation from in-control PMF $\boldsymbol{p}_0$.

Sample-based process monitoring

In quality-related applications, commonly *conforming category*, say s_0, much more frequent than *defect categories* $s_1, \ldots, s_d$. So $\boldsymbol{p}_0$ often close to one-point distribution $\boldsymbol{e}_0$ (low dispersion).

$\Rightarrow$ control charts monitoring categorical dispersion!

- **IQV chart** for nominal event series (Weiß, 2018b), **IOV chart** for ordinal ones (Bashkansky & Gadrich, 2011):

$$\mathrm{IQV}_r = \tfrac{d+1}{d}\Big(1 - \sum_{j=0}^{d} \frac{N_{r,j}^2}{n^2}\Big), \qquad \mathrm{IOV}_r = \tfrac{4}{d} \sum_{j=0}^{d-1} \frac{C_{r,j}}{n}\Big(1 - \frac{C_{r,j}}{n}\Big).$$

Sample-based process monitoring

. . . for nominal event series with

inherent memory for sensitivity towards small changes:

- Sample version of **log-LR CUSUM** by Ryan et al. (2011):

  $$\mathsf{C}_r \;=\; \max\{0,\, \ell R_r + \mathsf{C}_{r-1}\} \qquad \text{with } \ell R_t \;=\; \sum_{j=0}^{d} N_{r,j}\, \ln\left(\frac{p_{1,j}}{p_{0,j}}\right).$$

  Considers "weighted class count" (weights $\ln\big(p_{1,j}/p_{0,j}\big)$).

- Perry (2020) applies **EWMA approach**

  to weighted class counts with weights $1/(n\, p_{0,j})$,

  i. e., no anticipated out-of-control scenario.

Sample-based process monitoring

... for ordinal event series often use "weighted class counts":

- **Demerit charts** (Montgomery, 2009, Section 7.3.3): use demerit weights $0 = w_0 < w_1 < \ldots < w_d$ to reflect severity of defect categories

$$D_r \;=\; w_1\, N_{r,1} + \ldots + w_d\, N_{r,d}.$$

- Sometimes, weights by statistical reasoning, such as **log-LR CUSUM** by Steiner et al. (1996) considering latent-variable approach for out-of-control scenario.

Sample-based process monitoring

... for ordinal event series:

- **SOC chart** (simple ordinal categorical) by Li et al. (2014), motivated by latent-variable approach (logit model):

$$\mathrm{SOC}_r \;=\; \left|\sum_{j=0}^{d}(f_{0,j-1}+f_{0,j}-1)\,N_{r,j}\right| \quad \text{with } f_{0,-1}:=0,$$

where average cumulative proportions $\frac{1}{2}(f_{0,j-1}+f_{0,j})$ known as "ridits" (Agresti, 2010, p. 10).

Li et al. (2014) suggest to substitute raw frequencies $\boldsymbol{N}_r$ by EWMA frequencies: $\boldsymbol{N}_r^{(\lambda)} = \lambda\,\boldsymbol{N}_r + (1-\lambda)\,\boldsymbol{N}_{r-1}^{(\lambda)}$.

Sample-based process monitoring

. . . for ordinal event series:

- **ACD chart** (average cum. data) by Wang et al. (2018):

$$\mathsf{ACD}_r \;=\; n^{-1}\left(\boldsymbol{N}_r - n\,\boldsymbol{p}_0\right)^\top \mathbf{V}\left(\boldsymbol{N}_r - n\,\boldsymbol{p}_0\right) \qquad (\text{or } \boldsymbol{N}_r^{(\lambda)}),$$

quadratic-form statistics with weight matrix $\mathbf{V}$.

If $\mathbf{V}^{-1} = \mathrm{diag}(\boldsymbol{p}_0)$, then χ^2-**chart**.

Wang et al. (2018): $\mathbf{V} = \mathbf{L}^\top \mathrm{diag}(\boldsymbol{w})\,\mathbf{L}$ with weight vector $\boldsymbol{w} = \mathbf{1} = (1, \ldots, 1)^\top$ and triangular $\mathbf{L}$: lower (upper) triangle filled with 2 (0), main diagonal with 1. (. . .)

Sample-based process monitoring

... for ordinal event series:

- (...) Then, **ACD chart** has statistics

$$\mathrm{ACD}_r \;=\; n^{-1}\,\sum_{j=0}^{d} w_j\,\Big(C_{r,j-1}+C_{r,j}-n\,\big(f_{0,j-1}+f_{0,j}\big)\Big)^2,$$

  so again “ridits”.

- References on further charts in Weiß (2021).

  Application to two real-world data examples later in talk.

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Machine Learning Approaches for Event Sequence Monitoring

Sketch

General idea: Apply rule-based procedures from temporal data mining (Laxman & Sastry, 2006; Weiß, 2017) to quality-related processes.

This link was first studied by Göb (2006).

Most relevant: procedures of **episode mining**, where rules generated based on available categorical event sequence (“Phase-I data”), then applied to forecasting events in ongoing process (“Phase-II application”).
While control charts trigger alarm once limits violated, rule-based proc. require action once critical event predicted.

For illustration of **episode mining**,

see approaches by Mannila et al. (1997):

Aim: derive and apply rules such as

if "episode" $(x_{t-2}, x_{t-1}, x_t) = (s_0, s_1, s_0)$ observed,

then expect $X_{t+1} = s_2$ with some "confidence".

Notation: "$\boldsymbol{a} \Rightarrow \boldsymbol{b}$" with $\boldsymbol{a} = (s_0, s_1, s_0)$ and $\boldsymbol{b} = (s_0, s_1, s_0, s_2)$.

To prevent spurious correlation,

episodes must satisfy given **support** requirement:

"frequency" of episodes has to exceed threshold value supp_{min}.

Such episodes called **frequent episodes**.

For frequent episodes $\boldsymbol{a}, \boldsymbol{b}$ with $\boldsymbol{a}$ sub-episode of $\boldsymbol{b}$,
rule $\boldsymbol{a} \Rightarrow \boldsymbol{b}$ evaluated by **confidence** $\quad \text{conf}(\boldsymbol{a} \Rightarrow \boldsymbol{b}) = \dfrac{\text{supp}(\boldsymbol{b})}{\text{supp}(\boldsymbol{a})}$,
interpreted as predictive power of $\boldsymbol{a} \Rightarrow \boldsymbol{b}$.

Permitted episodes not only single tuples from $\mathcal{S} \times \mathcal{S}^2 \times \ldots$,
but also sets of tuples possible, using operators like

- wildcard "*" for arbitrary symbol from $\mathcal{S}$, or
- parallel episode "[. . .]" for arbitrary ordering of symbols.

Example: $\quad \big(s_0, *\big)$ corresponds to set $\big\{(s_0, s_0), \ldots, (s_0, s_d)\big\}$,
and $\big(s_0, [s_1, s_2]\big)$ to $\big\{(s_0, s_1, s_2), (s_0, s_2, s_1)\big\}$.

Highly efficient algorithms for frequent episodes mining by applying famous **Apriori principle** of Agrawal & Srikant (1994):

Episode $\boldsymbol{a}$ can only be frequent (i. e., $\text{supp}(\boldsymbol{a}) > \text{supp}_{\text{min}}$) if *all* its sub-episodes frequent as well.

$\Rightarrow$ **bottom-up approach:** given set $\mathcal{F}_{k-1}$ of frequent episodes of length $k-1$, set $\mathcal{C}_k$ with candidate episodes of length k constructed by combining episodes from $\mathcal{F}_{k-1}$.
Then, their support determined to get $\mathcal{F}_k \subseteq \mathcal{C}_k$.

Further details and illustrative example in Weiß (2021).

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH
STAT

A Nominal Event Sequence on Paint Defects

Real Application

Real Application: Nominal Event Sequence

Manufacturing of ceiling fan covers (Mukhopadhyay, 2008) with possible paint defects (among $d = 6$ defect categories):

$s_0 =$ "no defect",

$s_1 =$ "poor covering", $s_2 =$ "overflow", $s_3 =$ "patty defect",

$s_4 =$ "bubbles", $s_5 =$ "paint defect", $s_6 =$ "buffing".

Like in Weiß (2018b), assume in-control PMF

$$\boldsymbol{p}_0 = (0.769, 0.081, 0.059, 0.021, 0.023, 0.022, 0.025)^\top,$$

i. e., if rth sample of size n_r, then $\boldsymbol{N}_r \sim \text{Mult}(n_r, \boldsymbol{p}_0)$.

If just "defect — yes or no", then rth defect count $Y_r \sim \text{Bin}(n_r, \pi_0)$ with $\pi_0 = p_{0,1} + \ldots + p_{0,d} = 0.231$.

Table 1 of Mukhopadhyay (2008): 24 samples with (heavily) deviating sample sizes, $n_1, \ldots, n_{24}$ vary between 20 and 404.

For nominal events ($\boldsymbol{N}_r$), we use χ^2-chart and IQV chart, while for aggregated defect counts (Y_r), we use np-chart.

Chart design with target in-control ARL as $\text{ARL}_0 = 370.4$.

For np-chart, as for any Shewhart chart, $\text{ARL} = 1/P\big(Y \notin [l;u]\big)$, so control limits as $\alpha/2$- and $(1-\alpha/2)$-quantile of $\text{Bin}(n_r, \pi_0)$, where $\alpha = 1/370.4 \approx 0.00270$.

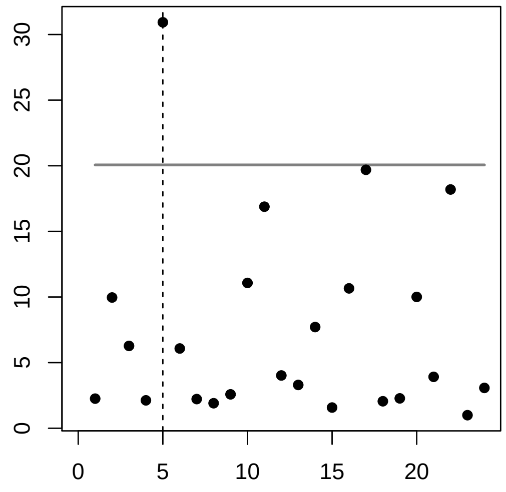
For X_r^2 and IQV_r, we use asymptotic distributions.

Real Application: Nominal Event Sequence

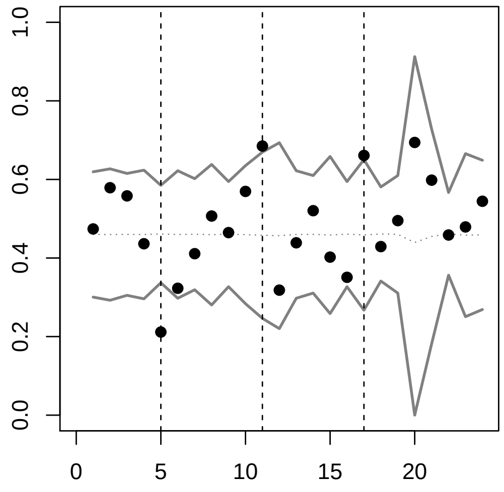
"Probability limits" $[l_r; u_r]$ vary according to sample sizes n_r,
also target ARL not met exactly because of discreteness.

For example, sample sizes $n_1 = 176$, $n_2 = 160$, ...
lead to $[l_1; u_1] = [25; 58]$, $[l_2; u_2] = [22; 54]$, ...
with individual in-control ARL values $444.4, 526.6, \ldots$

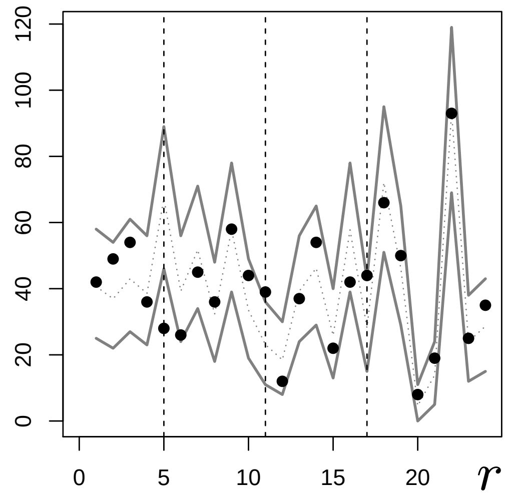
χ^2-chart



IQV chart



np-chart



Real Application: Nominal Event Sequence

χ^2-chart: alarm for 5th sample, but why? $\rightarrow$ "black box"

IQV chart: violation of lower limit for $r = 5$,
so dispersion decreased compared to $\boldsymbol{p}_0$,
so $\boldsymbol{N}_5/n_5$ approaches one-point distrib. in $s_0 =$ "no defect",
i. e., quality improvement (or problems in quality evaluation).

IQV chart (also np-chart): further alarms at $r = 11, 17$
by (slight) violation of upper limit,
so quality deterioration this time.

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

An Ordinal Event Sequence regarding Flash on Toothbrush Heads

Real Application

Real Application: Ordinal Event Sequence

Manufacturing of electric toothbrushes (Li et al., 2014).

In one production step, two parts of brush head welded together.

Excess plastic (“flash”) might occur, could injure users.

Thus, important quality characteristic:

extent of flash, with the $d+1=4$ ordinal levels

$s_0 =$ “slight”, $s_1 =$ “small”, $s_2 =$ “medium”, $s_3 =$ “large”.

In Phase-I analysis, Li et al. (2014) identify

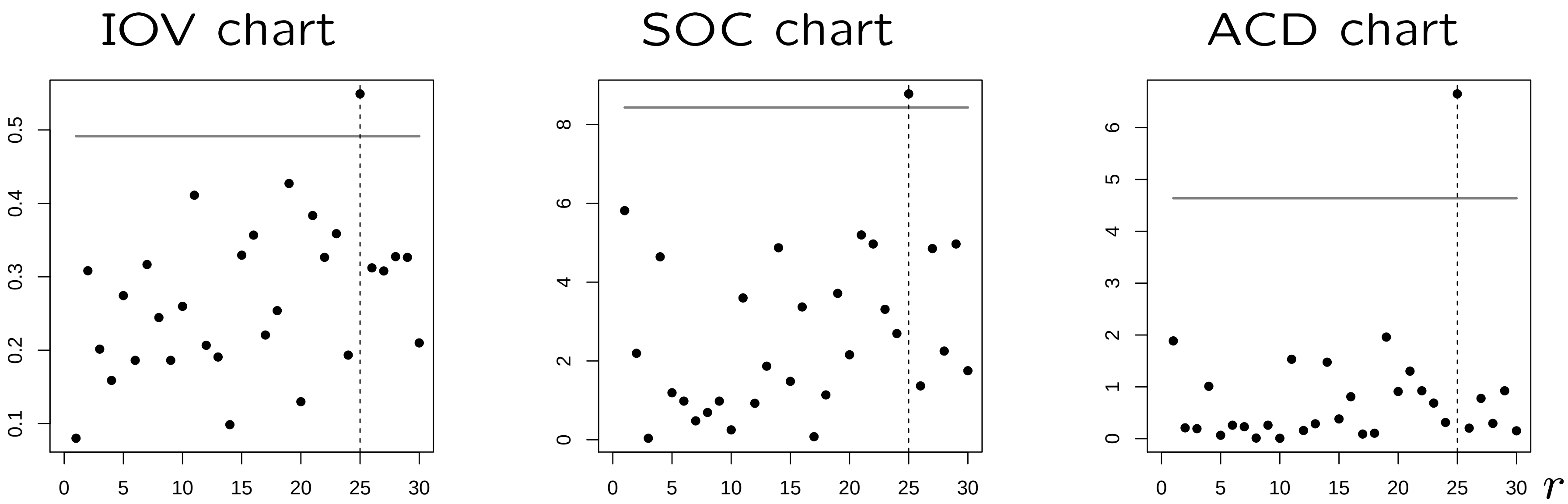
$\boldsymbol{p}_0 = (0.8631, 0.0804, 0.0357, 0.0208)^\top$, now for chart design.

For Phase-II application, 30 samples of unique size $n = 64$.

We apply (upper-sided) IOV, SOC, and ACD chart.

First, charts without EWMA smoothing of frequencies $\boldsymbol{N}_r$.

Upper control limits (target $\text{ARL}_0 = 370.4$) by simulations, leading to 0.4915 for IOV chart, 8.432 for SOC chart, and 4.638 for ACD chart.

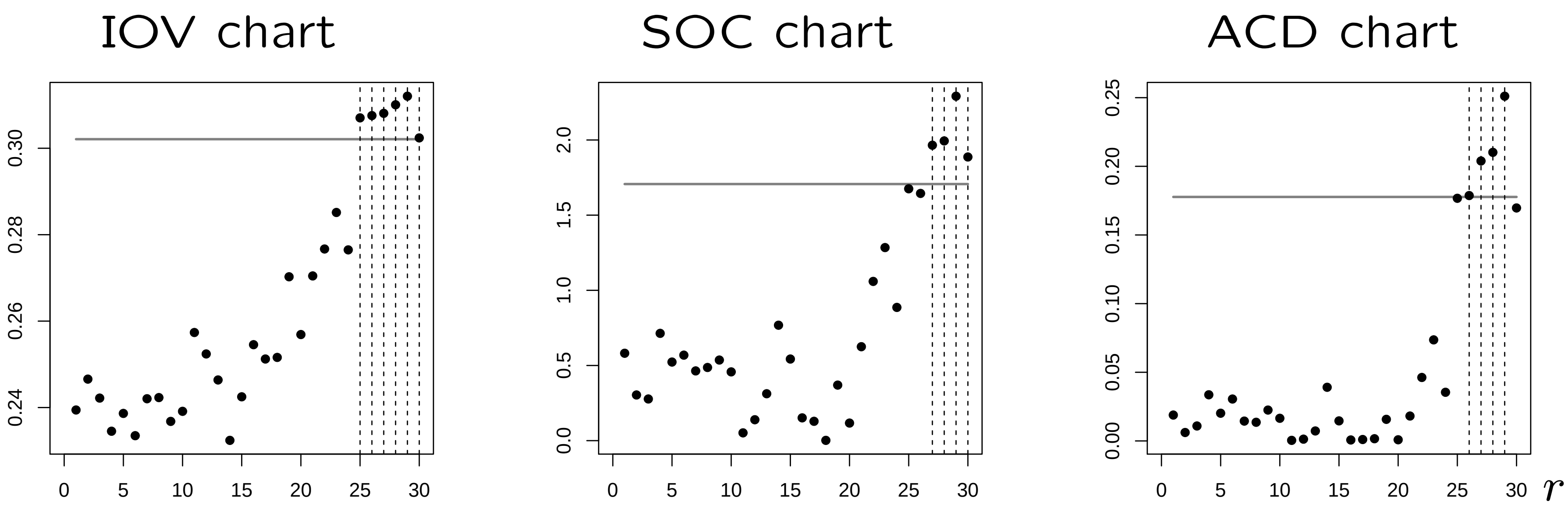


Alarm for 25th sample, quality deterioration.

Real Application: Ordinal Event Sequence

Next, additional EWMA smoothing, $\boldsymbol{N}_r^{(\lambda)} = \lambda\,\boldsymbol{N}_r + (1-\lambda)\,\boldsymbol{N}_{r-1}^{(\lambda)}$ with $\boldsymbol{N}_0^{(\lambda)} = n\,\boldsymbol{p}_0$ and $\lambda = 0.1$ (also see Li et al., 2014).

More narrow control limits, 0.3021 for IOV chart, 1.707 for SOC chart, 0.1777 for ACD chart.



First alarms at $r = 25$ (IOV), $r = 27$ (SOC), $r = 26$ (ACD).

Overall conclusion: problem with 25th sample.

This sudden change rather strong, so immediately detected by Shewhart charts. Some EWMA charts react with delay as EWMA charts more well-suited for persistent changes.

In fact, $\boldsymbol{N}_{25}/n \approx (0.7344, 0.1094, 0.06250.0938)^{\top}$;

so compared to $\boldsymbol{p}_0 = (0.8631, 0.0804, 0.0357, 0.0208)^{\top}$,

conforming probability for $s_0 =$ "slight" notably reduced,

while probability for worst state $s_3 =$ "large" increased.

So $\boldsymbol{N}_{25}/n$ moved towards extreme two-point distribution,

thus good performance of IOV charts.

Conclusions & Outlook

- Monitoring of categorical event series demanding task.
  During Phase I, appropriate methods for time series analysis
  and feasible stochastic models needed,
  while chart design for Phase-II application
  suffers from discreteness problems.

- **Future research:**
  Control charts for *serially dependent* categorical event series.
  In particular, tailored methods for ordinal time-series data,
  because particularly relevant for real applications.

Thank You

for Your Interest!

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Christian H. Weiß

Department of Mathematics & Statistics

Helmut Schmidt University, Hamburg

`weissc@hsu-hh.de`

Weiß (2021) On approaches for monitoring categorical event series. In K.P. Tran (ed.): *Control Charts and Machine Learning for Anomaly Detection in Manufacturing*, Springer Series in Reliability Engineering; 105–129.

Agrawal & Srikant (1994) Fast algorithms for ... *Proc. ICVLD*, 487–499.

Agresti (2010) *Analysis of Ordinal Categorical Data*. 2nd ed., Wiley.

Bashkansky & Gadrich (2011) Statistical quality ... *ASMBI* **27**, 586–599.

Duncan (1950) A chi-square chart for ... *IQC* **7**, 11–15.

Göb (2006) Data mining and statistical control ... *Frontiers* **8**, 285–308.

Höhle (2010) Online change-point detection ... Physica, 377–397.

Jacobs & Lewis (1983) Stationary discrete autoregr. . . *JTSA* **4**, 19–36.

Laxman & Sastry (2006) A survey of ... *Sādhanā* **31**, 173–198.

(...)

Literature

(. . .)

Li et al. (2014) A simple categorical chart . . . *IJPR* **52**, 550–562.

Li et al. (2018) Monitoring serially dependent . . . *IISET* **50**, 596–605.

Mannila et al. (1997) Discovery of frequent episodes . . . *DMKD* **1**, 259–289.

Montgomery (2009) *Introduction to Statistical Quality . . .* 6th ed., Wiley.

Mukhopadhyay (2008) Multivariate attribute control . . . *JAS* **35**, 421–429.

Page (1954) Continuous inspection schemes. *Biometrika* **41**, 100–115.

Perry (2020) An EWMA control chart . . . *JQT* **52**, 182–197.

Rakitzis et al. (2017) Control charts for . . . *ASMBI* **33**, 733–749.

Roberts (1959) Control chart tests . . . *Technometrics* **1**, 239–250.

Ryan et al. (2011) Methods for monitoring . . . *JQT* **43**, 237–248.

(. . .)

Literature

(. . .)

Steiner et al. (1996) Grouped data-sequential . . . *Technom.* **38**, 230–237.

Wang et al. (2018) A univariate procedure for . . . *CSSC* **47**, 115–128.

Weiß (2017) Association rule mining. *Wiley StatsRef Online*.

Weiß (2018a) *An Introduction to Discrete-Valued Time Series*. Wiley.

Weiß (2018b) Control charts for time-dependent . . . *Frontiers* **12**, 211–231.

Weiß (2020) Distance-based analysis . . . *JASA* **115**, 1189–1200.

Ye et al. (2002) Multivariate statistical analysis . . . *IEEETC* **51**, 810–820.